Short-Range Probabilistic Quantitative Precipitation Forecasts over the Southwest United States by the RSM Ensemble System

HUILING YUAN*

Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California

STEVEN L. MULLEN

Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona

XIAOGANG GAO AND SOROOSH SOROOSHIAN

Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California

JUN DU AND HANN-MING HENRY JUANG

Environmental Modeling Center, National Centers for Environmental Prediction, Washington, D.C.

(Manuscript received 12 May 2006, in final form 14 August 2006)

ABSTRACT

The National Centers for Environmental Prediction (NCEP) Regional Spectral Model (RSM) is used to produce twice-daily (0000 and 1200 UTC), high-resolution ensemble forecasts to 24 h. The forecasts are performed at an equivalent horizontal grid spacing of 12 km for the period 1 November 2002 to 31 March 2003 over the southwest United States. The performance of 6-h accumulated precipitation is assessed for 32 U.S. Geological Survey hydrologic catchments. Multiple accuracy and skill measures are used to evaluate probabilistic quantitative precipitation forecasts. NCEP stage-IV precipitation analyses are used as "truth," with verification performed on the stage-IV 4-km grid. The RSM ensemble exhibits a ubiquitous wet bias. The bias manifests itself in areal coverage, frequency of occurrence, and total accumulated precipitation over every region and during every 6-h period. The biases become particularly acute starting with the 1800–0000 UTC interval, which leads to a spurious diurnal cycle and the 1200 UTC cycle being more adversely affected than the 0000 UTC cycle. Forecast quality and value exhibit marked variability over different hydrologic regions. The forecasts are highly skillful along coastal California and the windward slopes of the Sierra Nevada Mountains, but they generally lack skill over the Great Basin and the Colorado basin except over mountain peaks. The RSM ensemble is able to discriminate precipitation events and provide useful guidance to a wide range of users over most regions of California, which suggests that mitigation of the conditional biases through statistical postprocessing would produce major improvements in skill.

1. Introduction

Cool season precipitation plays a central role in determining snowpack, runoff, and streamflow over the semiarid southwest United States (Serreze et al. 1999). Its importance to the hydrology of the region inspired Yuan et al. (2005a) to examine the performance of a 12-km version of the National Centers for Environmental Prediction (NCEP) Regional Spectral Model (RSM; Juang and Kanamitsu 1994) ensemble system for 24-h probabilistic precipitation forecasts over the southwest United States during the 2002/03 cool season. The RSM ensemble showed an overall wet bias. Forecast skill possessed large spatial variability over the region, however, with the highest skill over the coastal areas of California and windward of major mountain barriers

* Current affiliation: NOAA/Earth System Research Laboratory, Global Systems Division, and National Research Council Associate, Boulder, Colorado.

Corresponding author address: Huiling Yuan, NOAA/ESRL, R/GSD7, 325 Broadway, Boulder, CO 80305–3328.
E-mail: huiling.yuan@noaa.gov

DOI: 10.1175/MWR3373.1

and the lowest skill over the Great Basin and leeward slopes. Discrepancies also exist between the 0000 and 1200 UTC forecast cycles, with 0000 UTC runs being more skillful.

Hydrologists desire accurate estimates of quantitative precipitation amount and type at the finest possible spatial and temporal scale for flood and river flow forecasting models (Droegemeier et al. 2000). Operational streamflow models at the National Weather Service River Forecast Centers (RFCs) currently input 6-h precipitation totals (Charba et al. 2003), not 24-h accumulations. Thus, it is of interest to examine the accuracy of the 12-km RSM ensemble for 6-h intervals. The purpose of this paper is to explore the utility of 6-hourly probabilistic quantitative precipitation forecasts (PQPFs) for the same set of forecasts documented by Yuan et al. (2005a). Besides being of greater operational relevance, analysis of 6-h accumulation should help elucidate the differences at each 6-h forecast period between the skill of the 0000 and 1200 UTC forecasts and possibly reveal the reasons for the discrepancies.

Section 2 of this paper briefly describes the ensemble configuration and verification datasets. Section 3 describes the model performance of the 6-h PQPFs. Section 4 discusses the PQPFs and the economic values for different hydrologic regions. Section 5 presents the summary and the future research.

2. The RSM ensemble system and verification method

The NCEP RSM ensemble system is based on a 1997 version of the RSM (Juang and Kanamitsu 1994). The ensemble system uses regional breeding to generate its initial perturbations (Toth and Kalnay 1997; Du and Tracton 2001; Tracton and Du 2001), and forecasts from the NCEP global ensemble system (Toth and Kalnay 1993) to supply dispersive lateral boundary conditions. The RSM is one component of the NCEP Short-Range Ensemble Forecasting (SREF) system. The RSM configuration in the current SREF operational system is at an equivalent grid spacing of 45 km and runs to 87 forecast hours starting from 0900 and 2100 UTC initial conditions, with three grids covering the continental United States, Hawaii, and Alaska areas. The configuration includes one control run and two pairs of breeding members of the RSM forecasts (Du et al. 2006). Twice-daily forecasts to 24 h, starting from 0000 and 1200 UTC initial fields, were run for the 151 days from 1 November 2002 to 31 March 2003. The RSM has an equivalent grid spacing of 12 km, and the ensemble contains 11 members for each analysis cycle. The model domain (see Fig. 1 in Yuan et al. 2005a) covers two RFCs of the southwest United States, the California Nevada River Forecast Center (CNRFC) and the Colorado Basin River Forecast Center (CBRFC), and four U.S. Geological Survey (USGS) hydrologic unit regions: the Upper Colorado region, the Lower Colorado region, the Great Basin region, and the California region. The subregions are defined as in Fig. 1 and Table 1 by the USGS. Each catchment includes an area drained by a major river system.

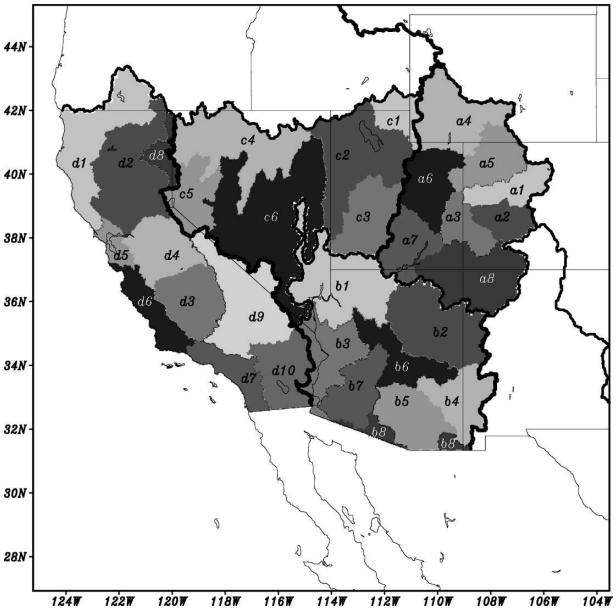

FIG. 1. Four USGS hydrologic regions (shaded area, i.e., the CNRFC and the CBRFC) and 32 catchments: the Upper Colorado region (labeled a), the Lower Colorado region (labeled b), the Great Basin region (labeled c), and the California region (labeled d). The list of names of the watersheds is given in Table 1.

Several verification measures, along with estimates of the confidence bounds, are used to evaluate the 6-h PQPFs. NCEP stage-IV precipitation analyses on a 4-km grid (hereafter termed stage IV) are used for "truth," with bilinear interpolation used to map the 12-km precipitation forecasts onto the stage-IV pixels. The 4-km comparison is used in order to maintain the stage-IV peak rainfall and maximize the study's applicability to high-resolution local weather forecasts and hydrological models. Summaries of domain-averaged skill are computed from local skill estimates before spatial averaging (Hamill and Juras 2007). While we assume an analysis as truth, the reader should keep in mind that skill scores can be highly dependent on the dataset used for verification, especially for precipitation forecasts (Yuan et al. 2005a).

For further details on the ensemble configuration and verification procedures not discussed above see Yuan et al. (2005a).

TABLE 1. The list of names for four USGS hydrologic regions and 32 catchments in Fig. 1.

a	Upper Colorado region
a1	Colorado Headwaters
a2	Gunnison
a3	Upper Colorado
a4	Great Divide
a5	White
a6	Lower Green
a7	Upper Colorado
a8	San Juan
b	Lower Colorado region
b1	Lower Colorado
b2	Little Colorado
b3	Lower Colorado
b4	Upper Gila
b5	Middle Gila
b6	Salt
b7	Lower Gila
b8	Sonora
c	Great Basin region
c1	Bear
c2	Great Salt Lake
c3	Escalante Desert
c4	Black Rock Desert
c5	Central Lahontan
c6	Central Nevada Desert basins
d	California region
d1	Klamath
d2	Sacramento
d3	Tulare
d4	San Joaquin
d5	San Francisco Bay
d6	Central California coastal
d7	Southern California coastal
d8	North Lahontan
d9	Northern Mojave
d10	Southern Mojave

3. Domain-averaged performance

Figure 2 shows the Brier skill score (BSS) and associated 90% confidence intervals (CIs). The CIs are estimated by nonparametric resampling based on a bootstrapping method (Efron and Tibshirani 1993; Hamill 1999), in which the BSS (and other verification metrics to follow) is calculated 10 000 times by repeatedly resampling the spatial statistic for a 6-h period of all verified samples. The results represent a spatially averaged value over the entire verification domain. Skill at the individual grid points is measured relative to sample climatology, the stage-IV data during the verification period, which is arguably a tougher competitor than long-term climatology. Sample climatology tends to capture the signal of seasonal climate trends and sample climatology frequency reflects occurrence variations during the study period, while long-term climatology and its frequency provide an average over a long-term period and neglect the anomaly during the sample data. Domain-averaged skill includes all pixels for which stage-IV data are available and an estimate of skill can be obtained. The figure reveals that the RSM ensemble only possesses skill during the first 6-h period, with confidence being highest for the 1200 UTC initial cycle. There are large discrepancies between the BSS for two forecast cycles during the second and third 6-h periods for all thresholds examined [1–20 mm $(6\ \text{h})^{-1}$]. The forecast skill for the 0000 UTC cycle monotonically decreases, while the BSS for the 1200 UTC cycle drops precipitously during the second 6-h period before flattening or slightly increasing afterward. Domain-averaged reliability diagrams (not shown) reveal that a persistent wet bias characterizes all forecast periods and thresholds, and that it is most severe during 1800–0000 UTC, or local afternoon hours. Consistency between the BSS and the bias from reliability diagrams indicates that the forecast bias is highly associated with the decrease of the forecast skill during 1800–0000 UTC.

The ranked probability skill score (RPSS) is an extension of the BSS concept to multicategorical forecasts (e.g., Wilks 2006). Figure 3 shows the RPSS based on five categories with boundaries of 1, 5, 10, and 20 mm $(6\ \text{h})^{-1}$, the same thresholds as in Fig. 2. Only the first 6-h period is skillful for both cycles; skill drops abruptly during the 1800–0000 UTC period for both cycles, consistent with the BSS results. The ranked probability score (RPS) can be decomposed (Hersbach 2000) into contributions from the reliability, resolution, and uncertainty terms. The reliability term is related to the conditional bias, the difference between the forecast probability and the observed probability conditioned on the forecast probability, and it is negatively oriented (smaller values are better). The resolution term measures the degree to which forecasts differ from climatology; it is positively oriented and is large for forecasts that differ greatly from climatology. The uncertainty term is related to the sample frequency of categorical occurrence and is independent of the forecast system. The RPSS is skillful relative to sample frequency when the resolution term exceeds the reliability term. The RPS decomposition (Fig. 4) indicates that the abrupt drop in skill results from a jump in the conditional bias during the 1800–0000 UTC time frame for both the 0000 and 1200 UTC cycles. The resolution term, which is related to the ability to discriminate events, steadily drops only ~20% over the 24-h period.

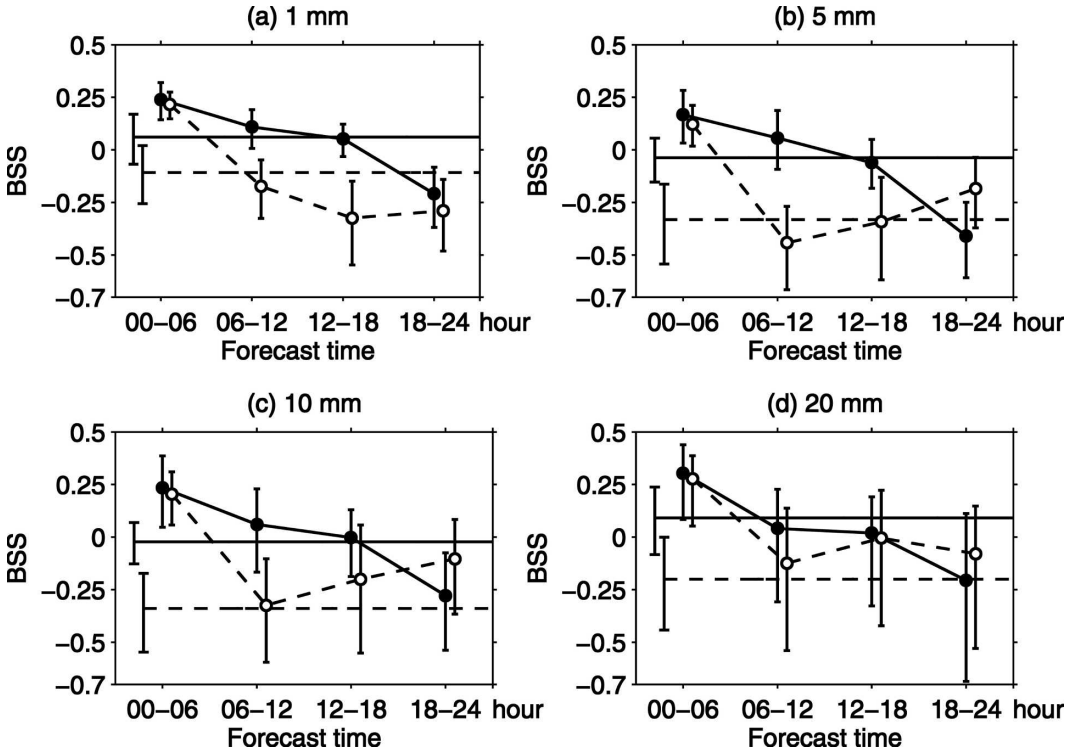

FIG. 2. The BSS of the 6-h PQPF of the CNRFC and the CBRFC districts for the 0000 UTC cycle (solid line with black circles) and 1200 UTC cycle (dashed line with white circles) at four thresholds [1, 5, 10, and 20 mm $(6\ h)^{-1}$]. Horizontal lines are the BSS of the 24-h PQPF for 0000 (solid) and 1200 UTC (dashed) at the same thresholds over two RFCs. Vertical bars indicate 90% CIs, computed from nonparametric resampling.

Rank histograms (Fig. 5) corroborate the differences between two analysis cycles. The enhanced population of the lowest ranks for all forecast projections indicates a persistent wet bias in frequency distribution. The bias is biggest, however, during the 1800–0000 UTC interval for both analysis cycles (the last 6-h period for the 0000 UTC cycle, the second for the 1200 UTC cycle) or 1000–1600 (1100–1700) Pacific (mountain) standard time. The occurrence of the maximum wet bias at the same local standard time (LST) means that the RSM does not properly capture the diurnal cycle of precipitation.

To examine the nature of the erroneous diurnal cycle in more detail, we compare the observed and modeled frequencies and accumulations of precipitation during the four 6-h periods for each analysis cycle. Figure 6 shows the occurrence frequency for a 5 mm $(6\text{-h})^{-1}$ threshold and seasonal time-mean accumulation for each 6-h period for total precipitation, along with convective and nonconvective contributions. Note that similar hatching on the bar charts corresponds to the same 6-h interval of LST.

Comparison of the initial 6-h forecasts to the corresponding 12–18-h forecasts *for the same LST* from the 12-h offset analysis cycle (cf. bars p1 with a3 and a1 with p3) reveals a "spinup" problem: the initial 6-h forecasts produce significantly lower (i.e., the mean values lie outside the range of confidence intervals for the comparison group) frequencies and amounts than the later forecast projections for the same LST. The spinup is

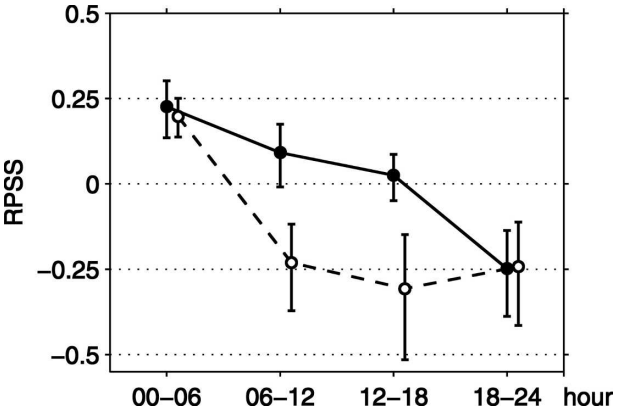

FIG. 3. The RPSS of 6-h PQPF for the CNRFC and the CBRFC districts using four thresholds [1, 5, 10, and 20 mm $(6\ h^{-1})$] for the 0000 (solid line with black circles) and 1200 UTC cycles (dashed line with white circles:). Vertical bars indicate 90% CIs.

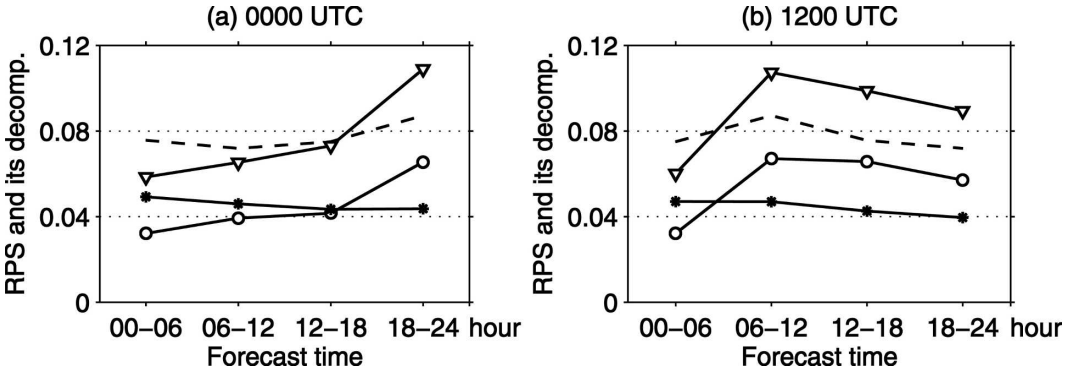

FIG. 4. Decomposition of the RPS of the CNRFC and the CBRFC districts for the (a) 0000 and (b) 1200 UTC forecasts. The dashed line is the uncertainty term; the solid line with open triangles is the RPS; the solid line with open circles is the reliability term; and the solid line with asterisks is the resolution term.

mostly confined to the model's nonconvective field, which produces between 5 and 10 times more precipitation than the direct contribution from the convective parameterization. The spinup problem persists to a somewhat lesser degree through the 6–12-h forecast of the 0000 UTC cycle (cf. bars p2 with a4), but not for the 1200 UTC cycle (cf. bars a2 with p4). It appears that the wet bias, which exists at all times and forecast cycles for frequency and accumulation (e.g., cf. o1 with p1 or a3), is especially large during the afternoon LST (bars a2 and p4). The convective component typically runs 5 times larger during the afternoon than the other periods, consistent with the earlier results of Hong and Leetmaa (1999) for a coarser-resolution version of the RSM. Their simulations also exhibited an early initiation of convection that resulted in large biases in the afternoon. They suggested that the trigger function in the cumulus parameterizations of the RSM should be investigated. Although the net contribution of the convective component to the total precipitation frequency is one-fifth that of the nonconvective contribution, we cannot rule out an indirect feedback mechanism affecting the nonconvective precipitation and playing a significant role in the total bias (Hong and Pan 1998). Hong and Pan (1998) examined cumulus parameterizations in the RSM and found that the feedback between convective and large-scale precipitation is nonlinear and involves highly complex interactions among all of the model physics and dynamics.

The frequency distributions for other thresholds and for afternoon convection exhibit similar behavior for thresholds between 1 and 20 mm $(6\ \text{h})^{-1}$. Moreover, it appears that a wet bias during the 2002/03 season is ubiquitous to other RSM configurations and the operational NCEP models. The 48-km RSM SREF and the Global Forecast System model (which supplies lateral boundary conditions for the RSM ensemble) possess a wet bias to 24 h (results not shown), but they are not as large as the RSM error.

The combination of spinup and an enhanced wet bias during the afternoon leads to an erroneous diurnal cycle. The stage-IV analyses indicate that the amplitude of diurnal cycle is between 5% and 10% of its daily mean value (Fig. 6b); whereas the amplitude for the RSM is much bigger, ~1/3 of its mean value. The maximum in total precipitation occurs during the 1800–0000 UTC period in the analyses and both forecast cycles, but the RSM minimum occurs 12 h out of phase in the 0000 UTC cycle, presumably because of the spinup. The deleterious impact of spinup during the initial 6-h period can be visualized by constructing a fictitious diurnal cycle for the RSM. Selecting the maximum value (or longest forecast projection) for the same LST from the 0000 or 1200 UTC cycle (i.e., an ordered grouping of a3, a4, p3, and p4) yields a diurnal cycle with minimal spinup and in better agreement with the corresponding observations (o1, o2, o3, and o4) in terms of relative amplitude (~20% of its mean value) and phasing (maximum during 1800–0000 UTC, minimum during 1200–1800 UTC). This hypothetical analysis suggests that use of a diabatic "hot start" initialization procedure (McGinley et al. 2000; Shaw et al. 2004) could improve certain aspects of the RSM diurnal cycle, but at the cost of amplifying the 0–6-h frequency bias.

4. Regional performance for watersheds of the Southwest

The RSM ensemble system exhibits significant spatial variations in 24-h skill over short distances, presumably in part because precipitation processes related to the complex surface heterogeneity of the southwest United States are not faithfully reproduced by the ensemble RSM analysis forecast system (Yuan et al.

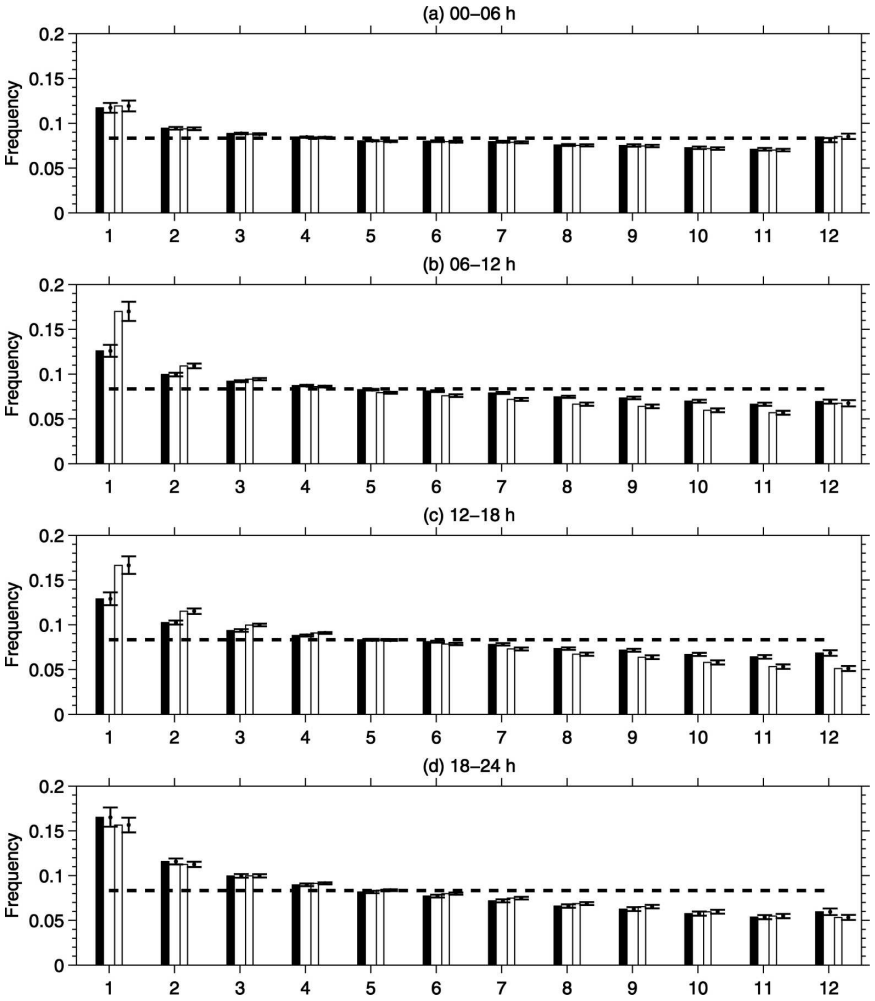

FIG. 5. Histograms of 6-h QPF for the 0000 UTC cycle (black bar) and 1200 UTC cycle (white bar) over the CNRFC and the CBRFC districts. The abscissa shows the rank of stage IV among the 11 forecast ensembles. The ordinate shows frequency. The horizontal line indicates the uniform rank distribution. The vertical bars indicate 90% CIs.

2005a). Thus, it is of interest to examine the spatial distribution of skill for 6-h accumulations, especially for the primary watersheds of the Southwest (Fig. 1) in view of ongoing efforts to couple atmospheric ensemble forecast systems and hydrologic runoff models. On the other hand, sample climatology frequency plays an important role on the verification skill (Yuan et al. 2005a; Hamill and Juras 2007). Lower sample climatology frequency could lead to lower skill scores in 6- and 24-h PQPFs with large uncertainties in the CIs, especially for extreme events over the drier areas.

Yuan et al. (2005a) compared 24-h PQPFs for the four large USGS hydrologic regions of the Southwest (Fig. 1). It is of interest to assess a more operationally relevant issue, the performance of 6-h RSM PQPFs for the 32 smaller catchments that make up the four Southwest USGS regions. Figure 7 presents the spatial distribution of the RPSS for the 0000 UTC forecasts for every stage-IV pixel. Skill varies widely across the model domain. RPSS values during every 6-h forecast interval range from ~0.5 (or greater) along the California coastal regions and the windward slopes of the Sierra Nevada Mountains and Mogollon Rim of Arizona, to below 0 during the initial 6-h period over vast regions of the Great Basin and the Four Corner States. Skillful regions can be generalized as being situated within 100 km of the Pacific coastline, upwind and along the crest of major mountain barriers, and unskillful ones as downwind of the major mountain barriers. Regions with positive skill steadily decrease with forecast projection in terms of area covered and skill level. By the last forecast period (1800–0000 UTC), skillful

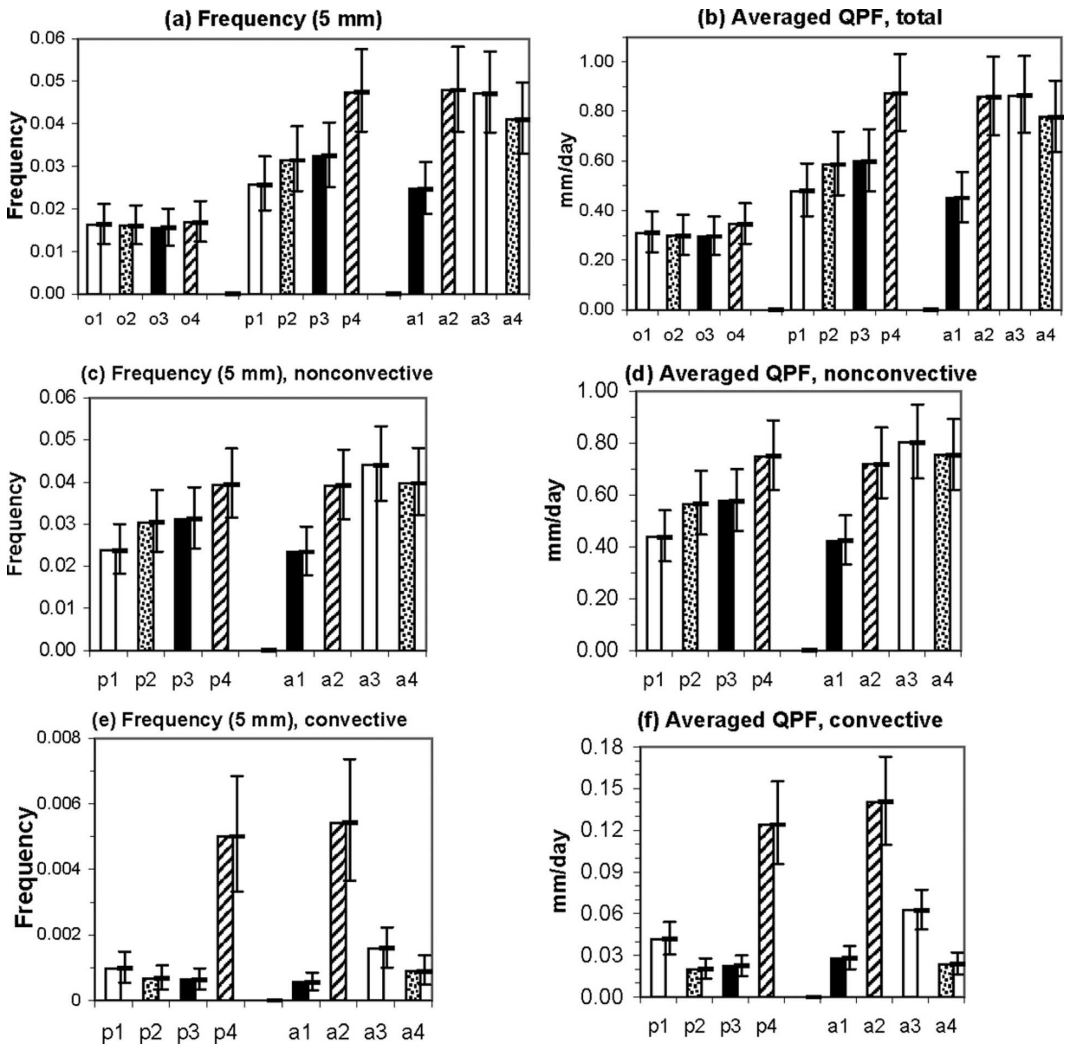

FIG. 6. Spatially averaged sample climatological frequencies and forecasted frequencies at the 5 mm $(6\ \text{h})^{-1}$ threshold and accumulated precipitation for the CNRFC and the CBRFC districts: (a), (b) 6-h total precipitation, (c), (d) 6-h nonconvective precipitation, and (e), (f) 6-h convective precipitation. Four 6-h periods: the 0000 UTC observations (o1, o2, o3, and o4); the 0000 UTC RSM forecasts (p1, p2, p3, and p4); and the 1200 UTC RSM forecasts (a1, a2, a3, and a4). The vertical bars indicate 90% CIs. The hatched bars indicate the cycles of the same 6-h UTC interval.

regions over the CNRFC and CBRFC districts are mostly confined to California, coastal Oregon, and the highest peaks of the interior Intermountain West. The 1200 UTC predictions exhibit similar tendencies but somewhat reduced skill starting during the 6–12-h forecast period (not shown), in accord with the domain-averaged results. Because the RPSS for catchments for the Upper Colorado basin and Great Basin generally lack skill, a subsequent discussion will emphasize those regions with appreciable skill.

The watershed with the best forecasts over the Lower Colorado basin is the Salt River basin (b6) of central Arizona, the primary water source for Phoenix, Arizona. Analysis of the RPSS and BSS, spatially averaged over the individual watersheds, shows that the b6 domain is skillful to 18 (6) h for the 0000 (1200) UTC forecasts (results not shown). The areas b3, b4, b5, b7, and b8 exhibit only minimal skill for the 1-mm threshold; for the 5-mm or higher thresholds, the entire Upper and Lower Colorado basins are unskillful. Most catchments over the Great Basin possess, at best, minimal skill, and typically no skill. The one exception is the c5 watershed that drains Lake Tahoe and supplies water to Reno and Carson City, Nevada. The 0000 UTC forecasts over c5 are skillful to 6 h for 10–20-mm thresholds and to 18 h for 5-mm and smaller thresholds,

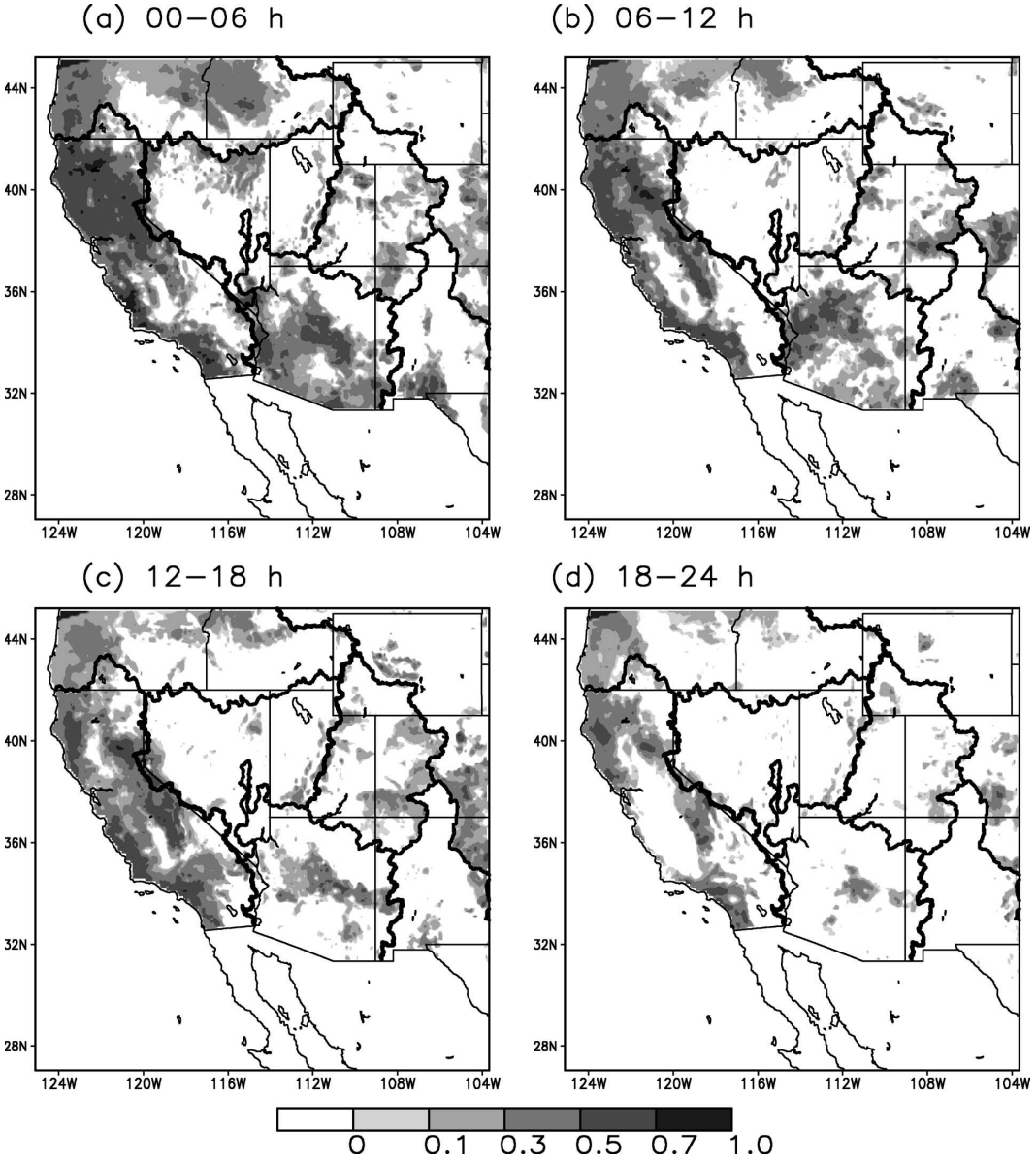

FIG. 7. Spatial distributions of the RPSS in each stage-IV pixel for 6-h precipitation during 151 days for the 0000 UTC cycle using four thresholds [1, 5, 10, and 20 mm $(6\ h)^{-1}$]. Boundaries for the USGS hydrologic regions are shown. Units are the same as in Fig. 1.

while the 1200 UTC forecasts show skill to 6 h for 1–15-mm thresholds.

California is the USGS hydrologic region with the best forecasts for 24-h accumulations (Yuan et al. 2005a), and as expected it is also the region with highest skill for 6-h accumulations over its 10 smaller catchments (Fig. 7). To illustrate commonalities and differences among the 10 catchments, we show area-averaged RPSS values for California catchments in Fig. 8. The north and south coastal areas (d1 and d7) are clearly skillful to 24 h in the sense that the CIs for every 6-h period are above the no skill line. The middle coastal areas (d5 and d6) and the Sacramento catchment (d2) are generally skillful to 24 h, although not every 6-h period has a 90% CI that lies above zero. Moreover, arid regions (d3, d4, d8, d9, and d10) that are located over interior central and southern California possess much lower skill levels than the coastal zones and Sacramento watershed. The Mojave districts (d9 and d10) especially stand out as lacking skill. There is a weak (though insignificant) tendency for the 0000 UTC forecasts to be slightly better than the 1200 UTC ones

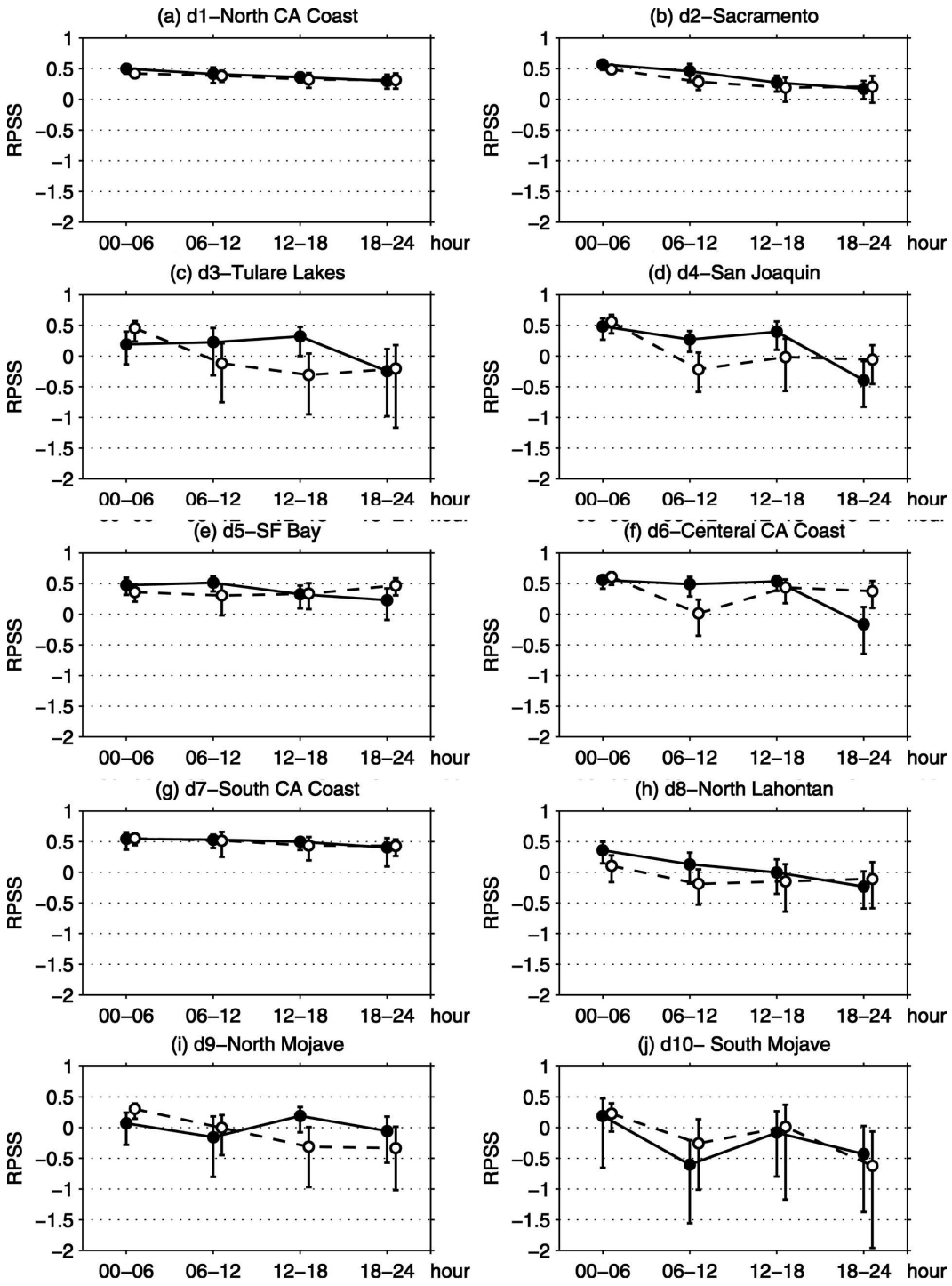

FIG. 8. The RPSS for each catchment over the California region for the 0000 UTC cycle (solid line with black circles) and 1200 UTC cycle (dashed line with white circles). The vertical bars indicate 90% CIs.

during the second and third 6-h periods, with the major exception being the Mojave Desert, which has a reverse signal.

The verification scores discussed so far are sensitive to biases because they are conditioned on the forecasts (i.e., if an event is forecast, what was observed?). The relative operating characteristic (ROC) curve stratifies forecasts based on observations (i.e., if an event is ob-

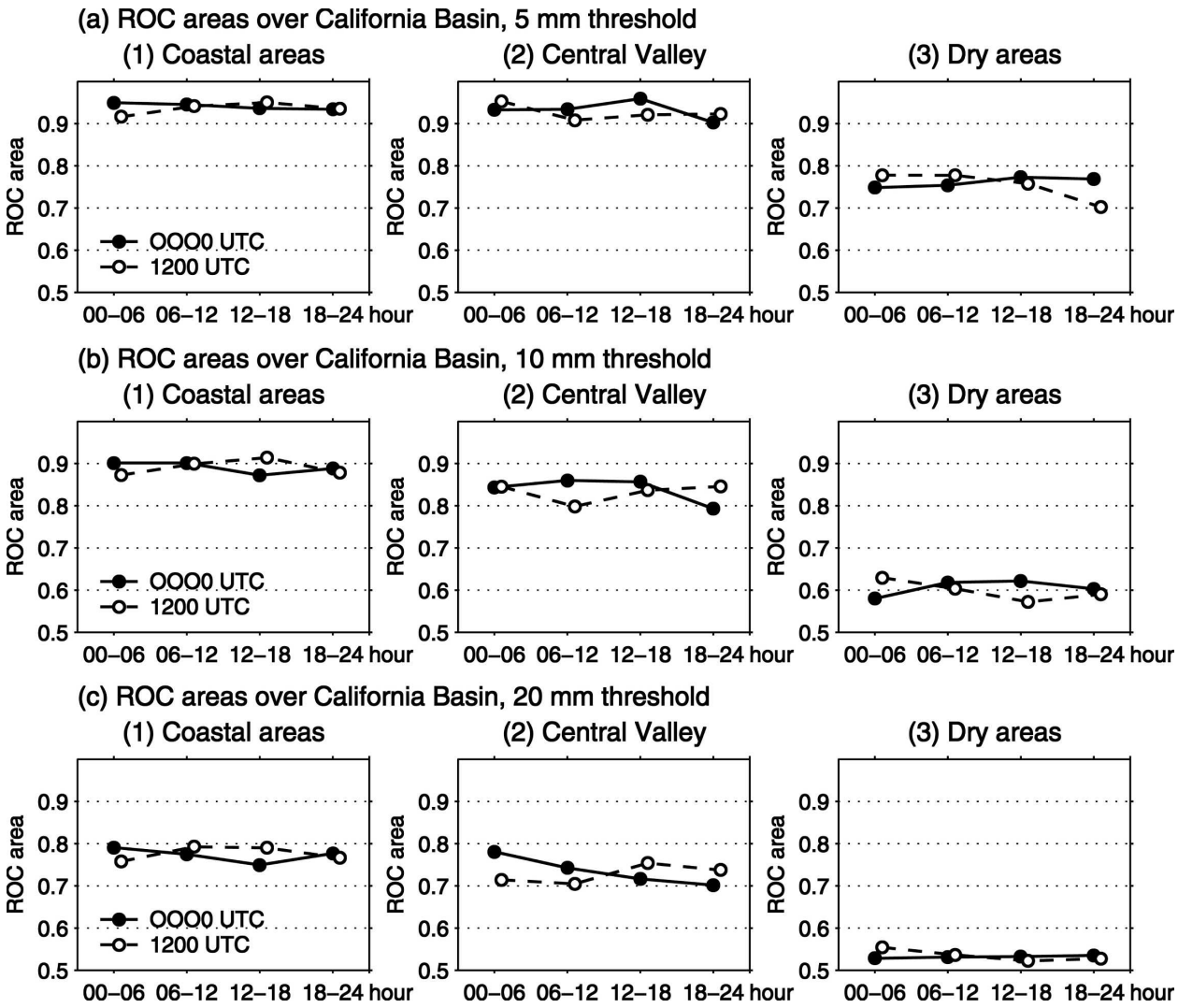

FIG. 9. Area under the ROC for three subgroups of catchments. (a1), (b1), (c1) Coastal areas: d1, d5, d6, and d7; (a2), (b2), (c2) Central Valley: d2, d3, and d4; (a3), (b3), (c3) dry interior basin areas: d8, d9, d10 over the California region at three thresholds [5, 10, 20 mm $(6\ \text{h})^{-1}$] for the 0000 UTC cycle (solid line with black circles) and 1200 UTC cycle (dashed line with white circles).

served, what was forecast?), and in that manner it measures the ability to discriminate between dichotomous events (e.g., rain versus no rain). Because the ROC curve is conditioned on the observations (Mason 1982), it is *insensitive* to model bias (Jolliffe and Stephenson 2003). Hit rate and false alarm rate pairs for all resolved RSM probability levels (e.g., 1/11, 2/11, . . . , 11/11), and the associated ROC curve, were computed at each stage-IV pixel for thresholds up to 20 mm $(6\ \text{h})^{-1}$. The area under the ROC curve provides a scalar measure of discrimination ability. The area is equivalent to the probability that a randomly selected "no" event will have a lower forecast probability than a randomly selected "yes" event (Hand and Till 2001), and thus area can be considered a measure of the separation between the two distributions. A ROC area of 1.0 denotes a perfect forecast, while an area of 0.5 or less indicates no skill relative to a climatology forecast. A value of ~0.75 represents a one standard deviation separation of the two distribution means under the assumption of homoscedasticity (i.e., the assumption of the constant variance across subsets of the data).

Figure 9 shows spatially averaged ROC areas that are averaged from ROC areas at each stage-IV pixel where a ROC curve can be defined, consistent with the recommendations of Hamill and Juras (2007). Composites for California are shown for three broadly similar geographic regions: coastal areas (catchments d1, d5, d6, and d7 in Fig. 1), the Central Valley (d2, d3, and d4) that includes the western slopes of the Sierra Nevada

Mountains, and dry areas (d8, d9, and d10) that contain interior desert basins. We show composites as opposed to spatial distributions as a way to mitigate the deleterious impact of the small number of pixels with heavy rain events [≥5 mm $(6\ \text{h})^{-1}$] over the California interior. The ROC areas reveal large differences between the composites. The coastal and Central Valley composites possess consistently higher ROC areas than the dry composite for all thresholds (5, 10, and 20 mm). The RSM performance is particularly impressive along the coastal regions, with ROC areas exceeding 0.90 and 0.75 for 10- and 20-mm thresholds, respectively; areas for the Central Valley typically run ~10% smaller across all thresholds, and they only dip slightly below 0.75 for 20 mm. The RSM ensemble is only able to discriminate well thresholds of 5 mm over the dry interior regions, where RSM forecasts for the heaviest threshold are effectively no better than a sample climatology forecast.

Potential economic value (PEV) curves can be derived from compositing the ROC curves of Fig. 9 (Richardson 2000; Buizza 2001; Zhu et al. 2002). PEV curves provide the optimal probability threshold at which a user should take a precautionary action at some cost C to prevent a larger loss L. The PEV represents the best forecast value among all forecast probabilities and its value ranges from $-\infty$ to 1. A PEV level of 1 denotes a perfect forecast, and a positive value denotes more useful information than that available from a climatology forecast. A negative PEV denotes worse guidance than climatology. The model has no ability to discriminate events for C/L ratios less than the lowest nonzero probability (1/11) provided by ensemble forecasts. Based on users' C/L ratio and the expected PEV, users can maximize value by choosing the optimal forecast probability threshold at which to take action, such as deciding when to release a reservoir for water managers. On another hand, for a very low or negative PEV, users cannot derive significant benefit from the forecast system over that provided by climatology.

Figure 10, which gives the PEV over the three California regions for a 10 mm $(6\ \text{h})^{-1}$ threshold, reveals that the value for the RSM ensemble varies by region. The RSM forecast system could provide decision makers with C/L ratios between 0.1 and 0.5 useful guidance over coastal California and the Central Valley. Users with the same range of C/L over the drier interior regions could also benefit, but the advantage over climatologic guidance would be lower. As can be inferred from the ROC areas (Fig. 9), the PEV curves also vary by precipitation threshold. (Results are not shown for other thresholds.) Low thresholds typically have a wider range of positive C/L and a higher optimal PEV levels, but even thresholds up to 20 mm $(6\ \text{h})^{-1}$ show

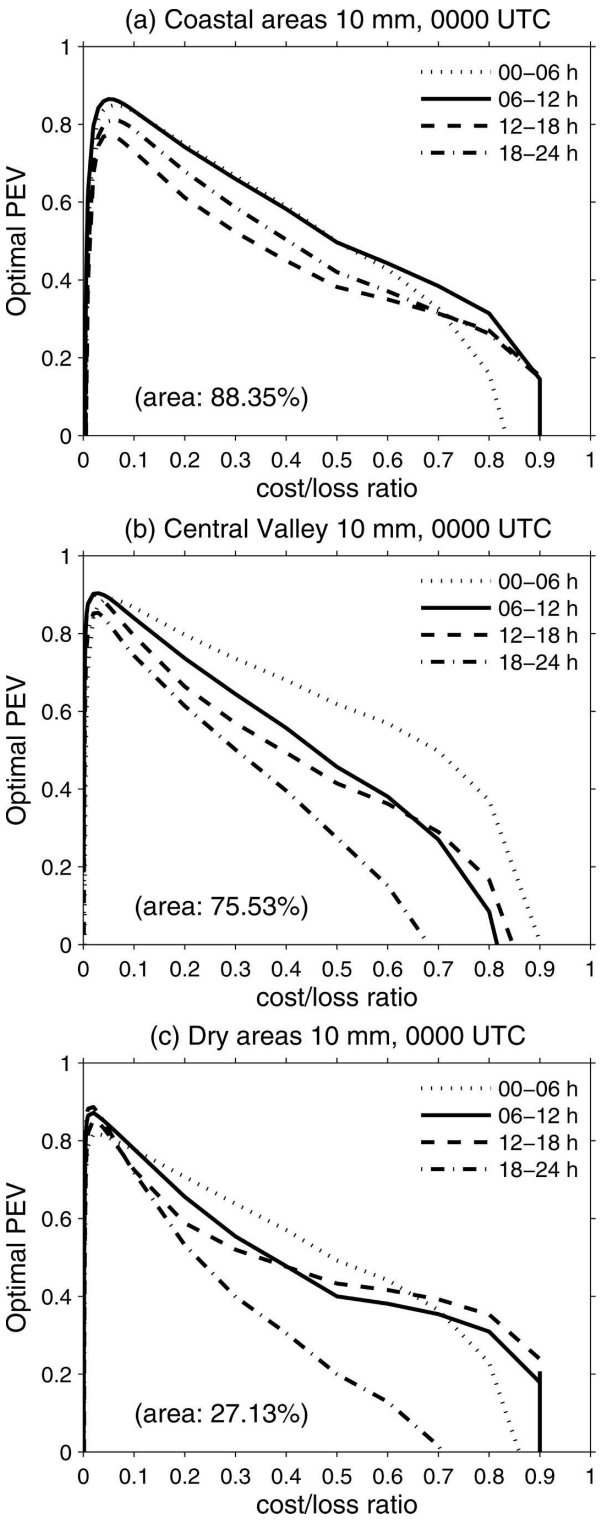

FIG. 10. PEV curves for 0000 UTC forecasts of a 10-mm threshold. Results are for a composite of stage-IV pixels with at least one observed event. Percentages in the lower left give the percent of pixels with at least one observed event for the subgroup: (a) coastal areas, (b) Central Valley, and (c) dry areas. Forecasts of 0–6 (dotted line), 6–12 (solid line), 12–18 (dashed line), and 18–24 h (dash–dotted line).

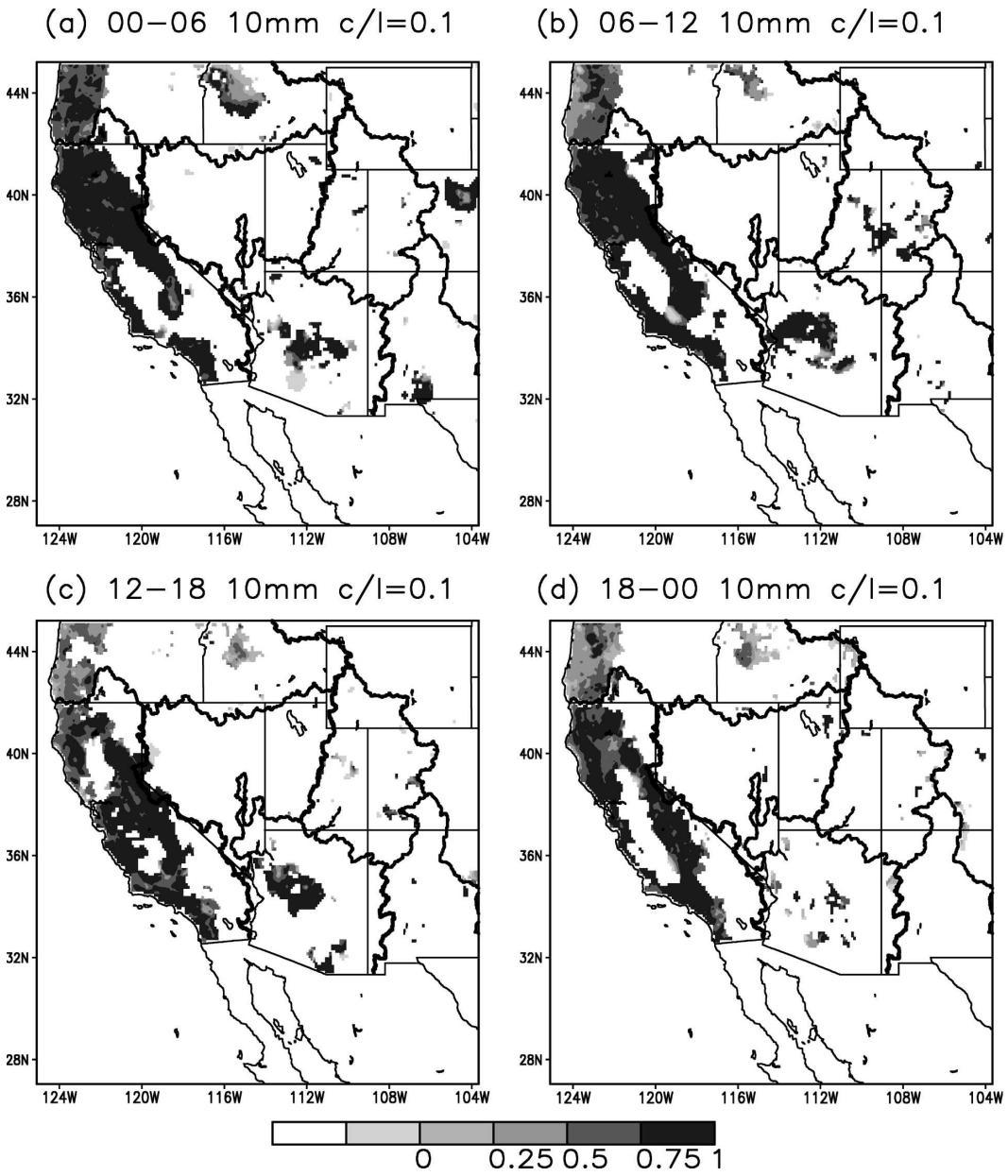

FIG. 11. PEVs at the 10-mm threshold ($C/L = 0.1$) for the 0000 UTC cycle forecasts of (a) 0–6, (b) 6–12, (c) 12–18, and (d) 18–24 h. Areas without shading indicate a sample climatology probability of zero. USGS hydrologic zones are as Fig. 1.

positive PEVs for a narrow range of C/L over coastal California and the Central Valley.

The spatial distribution of the maximum PEV values for the 10-mm threshold is shown in Fig. 11. Positive values are mostly confined to the coastal region, the wind slopes of the Sierra Nevada Mountains, and the mountains of Arizona. The PEV curves also exhibit a peculiar behavior: some 6-h accumulated precipitation forecasts at a longer range provide higher values than the shorter projections. For example, consider the western slopes of the Sierra Nevada Mountains, along 36°N. The 0–6-h forecast (Fig. 11a) shows PEVs less than 0.50, whereas the 6–12-h forecast (Fig. 11b) shows values greater than 0.75. Perhaps the most striking feature is the absence of 10-mm episodes over vast regions of the model domain during the 2002/03 season. Events did not occur over the San Joaquin Valley of California, or east of the crest of the Sierra Nevada Mountains and the Southern California mountains. The paucity attests to the challenge of acquiring a sufficient sample size for

heavy events of hydrological interest; these results for one season imply that a sample over many seasons is needed.

5. Summary

PQPFs for 6-h accumulations to 24 h from the NCEP RSM ensemble system were analyzed. The RSM was run twice daily at an equivalent grid spacing of 12 km during the 2002/03 cool season (1 November–31 March). The 4-km NCEP stage-IV precipitation analyses were used for verification, with the model output interpolated to the finer 4-km grid. The spatial variability and temporal evolution of the forecast skill were assessed for 32 catchments inside of the four USGS hydrologic regions of the southwest United States.

Analysis of the 6-h forecasts illuminated the underlying reason for the larger 24-h wet bias in the 1200 UTC forecasts compared with the 0000 UTC forecasts (Yuan et al. 2005a). The period 1800–0000 UTC marks the onset time of a large and sustained systematic error in both forecast cycles. Because the 1800–0000 UTC period occurs earlier in the 1200 UTC cycle, the overprediction of 24-h precipitation totals in the 1200 UTC runs is stronger than in the 0000 UTC cycle. This wet bias in the RSM ensemble is ubiquitous, affecting forecasts throughout the entire Southwest. Although the increase in parameterized convective precipitation frequency during the 1800–0000 UTC interval is 5 times bigger than during any of the other 6-h intervals, its net contribution to the total precipitation accumulation during every interval is much less than the contribution from grid-resolvable precipitation. The only period that does not contain a large wet bias is the initial 6-h forecast, which implies that spinup plays a significant role in mitigating the problem, that is, one error offsets another type of error. A more thorough analysis of the diurnal cycle would require a forecast length beyond 24 h, which is beyond the scope of this research. Even without such analysis, it is clear that decreasing the RSM grid spacing to 12 km did not appreciably improve the bias of coarser RSM simulations (Hong and Leetmaa 1999) and Global Spectral Model (GSM) forecasts.

The RSM ensemble exhibits large spatial variation in skill over the heterogeneous terrain of the southwest United States. PQPFs are more skillful along coastal areas and windward slopes of mountainous regions than over the lower elevations of arid inland regions. The underlying reasons for the regional variations in skill are not clear, and we dare not postulate potential mechanisms at this time. What is clear is that the heterogeneous Southwest, with its vast areas of sparse sampling and relatively few heavy precipitation events, poses serious challenges to improving ensemble performance.

Despite low skill or no skill over vast regions of the Southwest, the RSM ensemble is able to discriminate dichotomous precipitation events at the requisite hydrological temporal (6 h) and spatial (4 km) scales that are currently used by the NCEP operational runoff models (Charba et al. 2003). The RSM ensemble can provide useful guidance for a wide range of users, especially for many catchments over California where the ROC areas (range of 0.8–0.9) demonstrate that discriminating ability is quite high for thresholds up to 10 mm, and the areas are above 0.7 for even 20 mm. Even over the Great Basin and Upper Colorado River basin, ROC areas are sufficiently large ($\geq$0.7) and PEVs are sufficiently positive to indicate useful levels of discrimination and value for low thresholds (5 mm or smaller).

The ROC areas and PEV curves, metrics that are insensitive to conditional biases, also suggest that statistical postprocessing of the RSM PQPF fields could yield significant improvements in skill. Preliminary evaluation (Yuan et al. 2005b) shows that conditional biases, which are conditioned on the forecast probability categories, in the 24-h PQPFs from the 12-km RSM ensemble can be significantly reduced through calibration by an artificial neural network; there is no reason to believe calibration of the 6-h accumulations would behave differently. A single season of twice-daily forecasts is not a big enough sample to ensure a robust calibration for high thresholds (Hamill et al. 2006), however, so the extension of such high-resolution ensemble forecasts over multiple seasons is highly desirable.

We believe that the results from this study and the companion papers (Yuan et al. 2005a,b) support the notion of developing a high-resolution Rapid Update Cycle (RUC; Benjamin et al. 2004; Lee et al. 2006) ensemble system in conjunction with data assimilation systems, tailored to meet the needs of the hydrometeorological community (along with those of other end users). Such a system could provide timely, hourly updated guidance on the potential for heavy precipitation, severe storms, and flash flooding. Output from the atmospheric ensemble could drive a rapid update, hydrologic ensemble prediction system, an ensemble hydrometeorological analog to the RUC deterministic forecasts. We do believe that computing resources are not a major obstacle, as the technology currently exists to run such a coupled atmospheric–hydrologic system on reasonably priced clusters; the 12-km RSM ensemble was run in nearly real time on a 4-yr old cluster whose floating-point performance would cost approximately $50,000 or less in 2006 U.S. dollars.

Acknowledgments. The authors acknowledge the support of NASA EOS-IDS Grant NAG5-3460, NSF STC program (Agreement EAR-9876800). The second author (S. L. Mullen) also received partial support from ONR N00014-99-1-0181 and NSF Grants ATM-0135801 and ATM-0432232. Computing resources were obtained under support of ONR N00014-00-1-0613. Dr. Z. Toth and Mr. Y. Zhu provided the NCEP GSM ensemble data. The authors also thank the two anonymous reviewers for their insightful suggestions that led to many clarifications and major improvements in the manuscript. Ms. D. Hohnbaum and Ms. A. Reiser helped edit early versions of the paper.

REFERENCES

Benjamin, S. G., and Coauthors, 2004: An hourly assimilation—Forecast cycle: The RUC. *Mon. Wea. Rev.,* **132,** 495–518.

Buizza, R., 2001: Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev.,* **129,** 2329–2345.

Charba, J. P., D. W. Reynolds, B. E. McDonald, and G. M. Carter, 2003: Comparative verification of recent quantitative precipitation forecasts in the National Weather Service: A simple approach for scoring forecast accuracy. *Wea. Forecasting,* **18,** 161–183.

Droegemeier, K. K., and Coauthors, 2000: Hydrological aspects of weather prediction and flood warnings: Report of the ninth prospectus development team of the U.S. weather research program. *Bull. Amer. Meteor. Soc.,* **81,** 2665–2680.

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. on Mesoscale Processes,* Ft. Lauderdale, FL, Amer. Meteor. Soc., 355–356.

——, J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members. Preprints, *WMO Expert Team Meeting on Ensemble Prediction System,* Exeter, United Kingdom, 5 pp. [Available online at http://wwwt.emc.ncep.noaa.gov/mmb/SREF/reference.html.]

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman and Hall, 436 pp.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

——, and J. Juras, 2007: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.,* in press.

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important data set for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

Hand, D. J., and R. J. Till, 2001: A simple generalization of the area under the ROC curve for multiple class classification problems. *Mach. Learn.,* **45,** 171–186.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Hong, S.-Y., and H.-L. Pan, 1998: Convective trigger function for a mass-flux cumulus parameterization scheme. *Mon. Wea. Rev.,* **126,** 2599–2620.

——, and A. Leetmaa, 1999: An evaluation of the NCEP RSM for regional climate modeling. *J. Climate,* **12,** 592–609.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Wiley and Sons, 240 pp.

Juang, H.-M. H., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.,* **122,** 3–26.

Lee, M.-S., Y.-H. Kuo, D. M. Barker, and E. Lim, 2006: Incremental analysis updates initialization technique applied to 10-km MM5 and MM5 3DVAR. *Mon. Wea. Rev.,* **134,** 1389–1404.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

McGinley, J. A., S. Albers, D. Birkenheuer, B. Shaw, and P. Schultz, 2000: The LAPS water in all phases analysis: The approach and impacts on numerical prediction. Preprints, *Fifth Int. Symp. on Tropospheric Profiling,* Adelaide, Australia, Amer. Meteor. Soc., 133–135.

Richardson, D. S., 2000: Skill and economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

Serreze, M., M. Clark, R. Armstrong, D. McGinnis, and R. Pulwarty, 1999: Characteristics of western U.S. snowpack telemetry (SNOTEL) data. *Water Resour. Res.,* **35,** 2145–2160.

Shaw, B. L., S. Albers, D. Birkenheuer, J. Brown, J. McGinley, P. Schultz, J. Smart, and E. Szoke, 2004: Application of the Local Analysis and Prediction System (LAPS) diabatic initialization of mesoscale numerical weather prediction models for the IHOP-2002 field experiment. Preprints, *20th Conf. on Weather Analysis and Forecasting,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, P3.7.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Tracton, M. S., and J. Du, 2001: Short-Range Ensemble Forecasting (SREF) at the National Centers for Environmental Prediction. A report to WMO ensemble expert meeting, 10 pp. [Available online at http://wwwt.emc.ncep.noaa.gov/mmb/SREF/Tracton_Du.forWMO2001.doc.]

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2d ed. Elsevier Academic Press, 627 pp.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H. H. Juang, 2005a: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.,* **133,** 279–294.

——, ——, ——, and ——, 2005b: Calibration of probabilistic quantitative precipitation forecasts from the RSM ensemble forecasts over hydrologic regions. Preprints, *19th Conf. on Hydrology,* San Diego, CA, Amer. Meteor. Soc., CD-ROM, J3.4.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 73–83.